Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 129–137

www.elsevier.com/locate/specom

Under-determined reverberant audio source separation using Bayesian Non-negative Matrix Factorization

Sayeh Mirzaei[a,*], Hugo Van Hamme[a], Yaser Norouzi[b]

[a] Department of Electrical Engineering-ESAT, KULeuven, Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium
[b] Department of Electrical Engineering, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran

Received 28 May 2015; received in revised form 11 January 2016; accepted 13 January 2016
Available online 3 February 2016

Abstract

In this paper, we address the task of audio source separation for a stereo reverberant mixture of audio signals. We use a full-rank model for the spatial covariance matrix. Bayesian Non-negative Matrix Factorization(NMF)frameworks are introduced for factorizing the time-frequency variance matrix of each source into basis components and time activations. We also propose to incorporate the temporal dependencies in the Bayesian model through (1) recursively updating the prior hyperparameters or (2) applying a prior with Markov chain structure to favor the smoothness of the solution and we compare the performance of these two schemes. The EM algorithm is applied to derive the update relations of the unknown parameters. The separation performance improvement over the non-Bayesian standard NMF method as well as the conventional full-rank unconstrained method are investigated by calculating objective separation evaluation metrics.

Keywords: Blind Source Separation (BSS); Bayesian Non-negative Matrix Factorization(BNMF); Spatial covariance model; Temporal dependencies; Reverberant mixture.

1. Introduction

We often deal with a mixture of sounds coming from different acoustic sources. Separation of these audio signals and extracting the individual source signals is required in many applications including speaker diarization, meeting transcription systems, hearing aids, polyphonic music transcription, etc.

When no prior information of the sources or channel mixing system is available, the task is called Blind Source Separation (BSS). The multichannel mixture signal $\mathbf{x}(t) \in \mathbb{R}^M$ can be expressed as

$$\mathbf{x}(t) = \sum_{n=1}^{N} \mathbf{y}_n(t) \tag{1}$$

where $\mathbf{y}_n(t), \quad n = 1 \ldots N$ is the nth source spatial image vector over M channels. The mixing process consists of a linear time-invariant filtering of the source signals as:

$$\mathbf{y}_n(t) = \sum_{l=0}^{L-1} \mathbf{h}_n(l) s_n(t-l) \tag{2}$$

where $s_n(t)$ is the nth source signal, $\mathbf{h}_n(l) \in \mathbb{R}^M$ is the mixing filter vector which denotes the acoustic path from source n to the M microphones and L is the filter length. Most of the proposed BSS methods are based on the assumption that the mixing process at each frequency bin can be approximated by complex-valued multiplication:

$$\mathbf{Y}_n(f,t) \approx \mathbf{H}_n(f) S_n(f,t) \tag{3}$$

where $\mathbf{Y}_n(f, t)$ is the spatial image of source n in the Short Time Fourier Transform Domain(STFT) domain, $s_n(f, t)$ denotes the source STFT and $\mathbf{H}_n(f)$ specifies the Fourier transform of the mixing filter $\mathbf{h}_n(t)$.

If $S_n(f, t)$ is a zero-mean variable with variance $v_n(f, t)$, the covariance of $\mathbf{Y}_n(f, t)$ can be written as:

$$\mathbf{R}_{\mathbf{Y}_n}(f,t) = v_n(f,t)\mathbf{R}_n(f) \tag{4}$$

* Corresponding author. Tel.: +989126850714.
E-mail address: sayehm62@gmail.com (S. Mirzaei).

http://dx.doi.org/10.1016/j.specom.2016.01.003
0167-6393/

The assumption in (3) implies that the spatial covariance matrix of each source, $\mathbf{R}_n(f)$, has rank 1. Assuming the rank-1 model, BSS can be achieved using time-frequency (TF) masking techniques (Yilmaz and Rickard, 2004) or MAP estimation assuming sparse prior distributions (Winter et al., 2007), or modeling the source variances with Non-negative Matrix Factorization (NMF) (Févotte et al., 2009; Ozerov and Févotte, 2010). The rank 1 assumption is only valid when the filter length L is sufficiently small with respect to the STFT window length. This is violated in most realistic scenarios where reverberation exists. A full-rank spatial covariance matrix model is proposed in Duong et al. (2009) to provide better approximation in reverberant environments. The Maximum Likelihood (ML) solution is then found in an oracle context where both the spatial covariance matrix, $\mathbf{R}_n(f)$, and the scalar variance of the sources, $v_n(f, t)$, are known and also in a semi-blind context where the spatial covariance matrix is estimated from single-source training data. In Duong et al. (2010a), the EM algorithm was used for blindly estimating both of the above parameters. The source permutation problem which arises when the unknown parameters are independently estimated at each frequency bin, has also been solved in Duong et al. (2010a).

In Arberet et al. (2010), the source variances $v_n(f, t)$ are modeled by NMF and the EM algorithm is used for blindly estimating the parameters similar to what is done in Duong et al. (2010a). In Duong et al. (2010c), the use of a non-uniform TF representation on the auditory-motivated equivalent rectangular bandwidth(ERB) scale is investigated. It has been shown that this representation is beneficial for multichannel convolutive source separation provided that the full-rank covariance model is used. This has also been investigated in Burred and Sikora (2006) for instantaneous mixtures and (Vincent, 2006) for convolutive mixtures.

In Duong et al. (2010b), four specific covariance models including the rank-1 anechoic model, the rank-1 convolutive model, the full-rank direct+diffuse model and the full-rank unconstrained model are considered. A hierarchical clustering-based method is used to initialize the parameters. Also, a Direction of Arrival (DoA) based approach is proposed to align the order of the estimated sources across all frequency bins.

In Duong et al. (2013) some spatial location prior distributions consistent with the theory of statistical room acoustics are proposed for application to the spatial covariance matrices and EM algorithms are derived for Maximum a Posteriori (MAP) estimation. In Nikunen and Virtanen (2014), a spatial covariance matrix model is proposed which consists of a weighted sum of Direction of Arrival kernels. This covariance model is combined with the Complex NMF (CNMF) framework proposed in Sawada et al. (2013) and the update relations for finding the unknown parameters are subsequently derived.

In Arberet et al. (2010), the nth source variance matrix $\mathbf{V}_n(F \times T)$ consisting of the above variance elements, $v_n(f, t)$, is approximated as a product of two non-negative matrices $\mathbf{W}_n(F \times K)$ and $\mathbf{H}_n(K \times T)$ which specify the basis components and time activation matrices respectively. It is assumed that the number of the components K required for modeling each source is known in advance. However this may not be a suitable presumption when the goal is to blindly separate the individual source signals and there is no prior information about the source types. Here, we propose a Bayesian NMF framework to automatically infer the number of basis vectors for each source. In our first approach, we develop a Bayesian framework assuming that the time activation matrix elements H_n are random variables with a Gamma prior distribution. An EM algorithm is developed for deriving the update equations. The update relations given in Arberet et al. (2010) are replaced with the newly derived relations for the factors of the source variance matrices which are obtained through MAP estimation. We have also modeled the temporal dependencies through imposing constraints to the prior distribution of the temporal activations. A procedure inspired from Mohammadiha et al. (2012) is used for updating the scale parameters of the prior distributions of the time activations.

In the second approach, we favor the smoothness of the results through applying an inverse-Gamma chain prior distribution inspired from Févotte et al. (2009).

In Smaragdis et al. (2014), a comprehensive study of the NMF methods which model the temporal statistics is done. One flexible approach for considering the actual temporal dependencies is to impose constraints on the model activations (Essid and Févotte, 2013; Févotte, 2011; Févotte et al., 2009; Mohammadiha et al., 2013; 2012; Virtanen, 2007; Wilson et al., 2008). These approaches are called dynamic or smooth NMF. They differ by the used penalty term in non-probabilistic settings or by the choice of the observation model and prior structure in the Bayesian frameworks. In Virtanen (2007), temporal continuity and sparseness constraints are applied to the activation coefficients. Temporal continuity is favored by using a cost term which is the sum of squared differences between the activations in adjacent frames, and sparseness is favored by penalizing nonzero activations. A Non-negative Dynamical System (NDS) is introduced in Févotte et al. (2013) for modeling speech spectra. It can be regarded as an extension of NMF to support Markovian dynamics. Non-negativity preserving Gamma or inverse-Gamma Markov chain priors are considered in Févotte (2011); Févotte et al. (2009); Mohammadiha et al. (2013, 2012) and Markov random fields in Kim and Smaragdis (2013). In Nakano et al. (2010), the spectrogram of music signals is modeled as the combination of Markov-chained spectral patterns.

The approaches proposed in this paper can be regarded as Bayesian extensions of the method proposed in Arberet et al. (2010) accentuating the smoothness of the estimates. The Gamma prior model has been chosen for its effectiveness in modeling sparse parameters. Meanwhile, Gamma and inverse-Gamma prior distributions are preferred because we are going to model non-negative elements of the activation matrix, thus other sparse prior distributions such as Laplace cannot be useful here. The novel aspects of our proposed approaches can be summarized as follows:

- Bayesian NMF frameworks are proposed to factorize the source variance matrix in the full-rank model for the purpose of providing a more powerful model through applying suitable prior structures and avoiding over-fitted or under-fitted models.
- Temporal dependencies are taken into account via 1)Updating the scale hyperparameters of the time activation Gamma prior distributions; 2) Applying an inverse-Gamma Markov chain prior distribution.

The structure of the rest of the paper is as follows: We introduce our first proposed Bayesian approach which imposes a temporal continuity constraint through updating the scale hyperparameters of the Gamma prior in Section 2. We then explain the second Bayesian approach which uses an inverse-Gamma chain prior structure in Section 3. Various experimental settings and the performance results are presented in Section 4. Finally, we conclude in Section 5.

2. Bayesian NMF with Gamma prior model configuration

We admit the following generative model for the nth source spatial image $\mathbf{y_n}(f, t)$ (Arberet et al., 2010):

$$\mathbf{Y}_n(f,t) \sim \mathrm{N}_c\big(0, \mathbf{R}_{\mathbf{Y}_n}(f,t)\big) \tag{5}$$

where $\mathrm{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a proper complex Gaussian distribution:

$$\mathrm{N}_c(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \left|\pi \boldsymbol{\Sigma}^{-1}\right| \exp\left[-(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right] \tag{6}$$

The covariance matrix $\mathbf{R}_{\mathbf{Y}_n}(f,t)$ is given by (4) where the spatial covariance matrix $\mathbf{R}_n(f)$ is assumed as a full-rank unconstrained covariance model (Duong et al., 2010b) and $v_n(f, t)$ is the time-varying source variance which we approximate with the following factorized form:

$$v_n(t,f) = \sum_{k=1}^{K} w_{f,k}^{(n)} h_{k,t}^{(n)} \tag{7}$$

where $w_{f,k}^{(n)}, h_{k,t}^{(n)} \in \mathbb{R}^+$. Therefore, we can express the matrix $\mathbf{V}_n$ with entries $[\mathbf{V}_n]_{f,t} = v_n(f,t)$ as a product of two non-negative matrices $\mathbf{W}_n$ and $\mathbf{H}_n$ with entries $[\mathbf{W}_n]_{f,k} = w_{f,k}^{(n)}$ and $[\mathbf{H}_n]_{k,t} = h_{k,t}^{(n)}$:

$$\mathbf{V}_n = \mathbf{W}_n \mathbf{H}_n \tag{8}$$

Consequently, the source spatial image $\mathbf{Y}_n(f, t)$ can be written as the combination of K components:

$$\mathbf{Y}_n(f,t) = \sum_{k=1}^{K} \mathbf{Y}_{n,k}(f,t) \tag{9}$$

with the covariance matrix given by:

$$\mathbf{R}_{\mathbf{Y}_{n,k}}(f,t) = w_{f,k}^{(n)} h_{k,t}^{(n)} \mathbf{R}_n(f) \tag{10}$$

Here, we assume that the elements of the $\mathbf{H}_n$ matrix representing the time activations of the basis components in $\mathbf{W}_n$, are random variables with Gamma prior distribution, $Gamma(h_{k,t}^{(n)} | a_{k,t}^{(n)}, b_{k,t}^{(n)})$:

$$Gamma(h|a,b) = \frac{h^{(a-1)}}{b^a \Gamma(a)} \exp\left(\frac{-h}{b}\right) \tag{11}$$

The elements of $\mathbf{W}_n$ are assumed deterministic.

According to the above generative model and assuming that the source signals are independent, the mixture signal STFT, $\mathbf{X}(f, t)$ would be a zero-mean complex Gaussian vector with the following covariance matrix:

$$\mathbf{R}_{\mathbf{X}}(f,t) = \sum_{n=1}^{N} \mathbf{R}_{\mathbf{Y}_n}(f,t) \tag{12}$$

The aim is to estimate the spatial covariance matrices $\mathbf{R}_n(f)$ and the source variances $v_n(f, t)$ under the above generative model by maximizing the likelihood function. We opt for the EM algorithm similar to Arberet et al. (2010); Duong et al. (2010b) with the difference that we maximize the posterior probability while updating the $\mathbf{H}_n$ elements. Therefore, the E-step of the EM algorithm remains unchanged w.r.t what is given in Arberet et al. (2010). The following M-step update relations are obtained:

$$w_{f,k}^{(n)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\hat{v}_{n,k}(f,t)}{h_{k,t}^{(n)}} \tag{13}$$

$$h_{k,t}^{(n)} = \frac{a_{k,t}^{(n)} - (F+1) + \sqrt{\left(a_{k,t}^{(n)} - (F+1)\right)^2 + \frac{4 \sum_{f=1}^{F} \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}}}{b_{k,t}^{(n)}}}}{\frac{2}{b_{k,t}^{(n)}}} \tag{14}$$

with $\hat{v}_{n,k}(f,t) = \frac{1}{M} tr(\mathbf{R}_n^{-1}(f) \hat{\mathbf{R}}_{y_{n,k}}(f,t))$. F denotes the total number of frequency bins in the TF representation. $\hat{\mathbf{R}}_{y_{n,k}}(f,t)$ is updated according to Eq. (11) of the E-step given in Arberet et al. (2010) and is restated in (A.2). $\mathbf{R}_n(f)$ is updated as (Arberet et al., 2010):

$$\mathbf{R_n}(f) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{v_n(f,t)} \hat{\mathbf{R}}_{\mathbf{Y}_n}(f,t) \tag{15}$$

More explanation on the derivation of (14) can be found in Appendix A.1.

After the convergence of the EM algorithm, the STFT of the source spatial image is estimated using the Wiener estimator:

$$\hat{\mathbf{Y}}_n(f,t) = \mathbf{R}_{\mathbf{Y}_n}(f,t) (\mathbf{R}_{\mathbf{X}}(f,t))^{-1} \mathbf{X}(f,t) \tag{16}$$

The time-domain signals are simply derived through inverse STFT operation.

The proposed Bayesian framework has the advantage of avoiding overfitted or underfitted models. Meanwhile, inspired from Mohammadiha et al. (2012), we recursively update the prior hyperparameters over subsequent time frames as described in Section 2.1 to impose the temporal continuity constraint. This continuity normally exists for many audio signals and for speech in particular.

2.1. Imposing the temporal continuity constraint

To make the model fit better to the data, we take the temporal dependencies into account. The scale parameters $b_{k,t}^{(n)}$ of

the prior are recursively updated at each time frame based on the following smoothing relation (Mohammadiha et al., 2012):

$$b_{k,t}^{(n)} = \lambda \frac{h_{k,t-1}^{(n)}}{a_{k,t}^{(n)}} + (1-\lambda) b_{k,t-1}^{(n)} \tag{17}$$

where the λ parameter controls the smoothing level. The shape hyperparameter $a_{k,t}^{(n)}$ is taken fixed and time-invariant in our model.

2.2. Parameter initialization

The EM algorithm is very sensitive to initialization, thus here, similar to Arberet et al. (2010), we use the perturbed oracle initialization where the parameters $\mathbf{R}_n(f)$ and $v_n(f, t)$ are estimated form the original source spatial images as in Duong et al. (2009) and then perturbed with a high level additive noise (SNR of 5 dB). The parameters $w_{f,k}^{(n)}, h_{k,t}^{(n)}$ are then initialized according to the IS-NMF Bayesian generative model with Gamma prior. The update relations are derived in Appendix B. We have chosen IS divergence because it has been shown to perform more effectively for factorizing the power spectrogram (Févotte et al., 2009). Furthermore, IS divergence is scale-invariant, which means that the same relative weight is given to small and large values of the source variance matrix coefficients $\mathbf{V}_n(F \times T)$ in the cost function. This property can be regarded as a benefit since it is relevant to decomposition of audio spectra in the sense that a bad fit of the factorization for a low-power coefficient will cost as much as a bad fit for a higher-power coefficient.

3. Bayesian NMF with inverse-Gamma chain prior model configuration

In the second Bayesian framework, we assume the inverse-gamma prior distribution for $h_{k,t}^{(n)}$ parameters as follows:

$$p\left(h_{k,t}^{(n)} | h_{k,t-1}^{(n)}\right) = IG\left(h_{k,t}^{(n)} | \alpha, (\alpha+1) h_{k,t-1}^{(n)}\right)$$
$$IG(u|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} u^{-(\alpha+1)} \exp\left(\frac{-\beta}{u}\right), \quad u \geq 0 \tag{18}$$

The prior is constructed so that its mode is obtained for $h_{k,t}^{(n)} = h_{(k,t-1)}^{(n)}$. α is a shape parameter that controls the sharpness of the prior around its mode. A high value of α will increase the sharpness and thus accentuate the smoothness of $h_k^{(n)}$, while a low value of α will render the prior more diffuse and thus less constraining (Févotte et al., 2009). In the following, $h_{k,1}^{(n)}$ is assigned the scale-invariant Jeffreys noninformative prior $p(h_{k,1}^{(n)}) \propto \frac{1}{h_{k,1}^{(n)}}$. In this case, the update relations are the same as in the previous section except for the $h_{k,t}^{(n)}$ parameters which are obtained through maximizing the posterior distribution. The update relation is obtained as:

$$h_{k,t}^{(n)} = \frac{-\sqrt{p_1^2 - 4p_2p_0} - p_1}{2p_2} \tag{19}$$

Table 1
Coefficients of the order 2 polynomial to solve in order to update $h_{k,t}^{(n)}$ in Bayesian IS-NMF with an inverse-Gamma chain prior.

	p_2	p_1	p_0
$h_{k,1}^{(n)}$	$-\frac{(\alpha+1)}{h_{k,2}^{(n)}}$	$-(F-\alpha+1)$	$\sum_f \frac{\hat{v}_{n,k}(f,1)}{w_{f,k}^{(n)}}$
$h_{k,t}^{(n)}, \ 1<t<T$	$-\frac{(\alpha+1)}{h_{k,t+1}^{(n)}}$	$-(1+F)$	$\sum_f \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}} + (\alpha+1)h_{k,t-1}^{(n)}$
$h_{k,T}^{(n)}$	0	$-(\alpha+1+F)$	$\sum_f \frac{\hat{v}_{n,k}(f,T)}{w_{f,k}^{(n)}} + (\alpha+1)h_{k,T-1}^{(n)}$

Table 2
Coefficients of the order 2 polynomial to solve in order to initialize $h_{k,t}^{(n)}$ in Bayesian IS-NMF with an inverse-Gamma chain prior.

	p_2	p_1	p_0
$h_{k,1}^{(n)}$	$\frac{(\alpha+1)}{h_{k,2}^{(n)}}$	$F-\alpha+1$	$-F\hat{h}_{k,1}^{(n)}$
$h_{k,t}^{(n)}, \ 1<t<T$	$\frac{(\alpha+1)}{h_{k,t+1}^{(n)}}$	$1+F$	$-F\hat{h}_{k,t}^{(n)} - (\alpha+1)h_{k,t-1}^{(n)}$
$h_{k,T}^{(n)}$	0	$(\alpha+1+F)$	$-F\hat{h}_{k,T}^{(n)} - (\alpha+1)h_{k,T-1}^{(n)}$

where the values of p_0, p_1 and p_2 are given in Table 1. The derivation of the above expression can be found in Appendix B.

3.1. Parameter initialization

Again, we use the perturbed oracle to initialize the parameters $\mathbf{R}_n(f)$ and $v_n(f, t)$. The parameters $w_{f,k}^{(n)}$ and $h_{k,t}^{(n)}$ are then initialized according to the IS-NMF with inverse-Gamma chain prior generative model. The update equations can again be obtained through finding the MAP estimates of the parameters. The $w_{f,k}$ update relation is the same as (A.8) and the $h_{k,t}$ coefficient is updated according to the Equation 5.10 in Févotte et al. (2009):

$$h_{k,t}^{(n)} = \frac{\sqrt{p_1^2 - 4p_2p_0} - p_1}{2p_2} \tag{20}$$

where

$$v'_{k,f,t,n} = \left| \frac{w_{f,k}^{(n)} h_{k,t}^{(n)}}{\sum_l w_{f,l}^{(n)} h_{l,t}^{(n)}} \right|^2 v_{f,t}^{(n)} + \frac{w_{f,k}^{(n)} h_{k,t}^{(n)}}{\sum_l w_{f,l}^{(n)} h_{l,t}^{(n)}} \sum_{l \neq k} w_{f,l}^{(n)} h_{l,t}^{(n)}$$
$$\hat{h}_{k,t}^{(n)} = \frac{1}{F} \sum_{f=1}^{F} \frac{v'_{k,f,t,n}}{w_{f,k}^{(n)}} \tag{21}$$

The values of p_0, p_1 and p_2 are given in Table 2.

4. Experiments

Here, we consider the stereo case ($M = 2$). Constant values are chosen for the parameters $a_{k,t}^{(n)} = 2$ of the Gamma prior and $\alpha = 5$ of the inverse-Gamma distributions. In a non-blind setting where some training data is available, these parameters can be learned instead of taking fixed values. The initial value for $b_{k,t}^{(n)}$ is chosen equal to the mean value of the $v_{f,t}^{(n)}$ matrix coefficients over f and t. The smoothing parameter λ is 0.5.

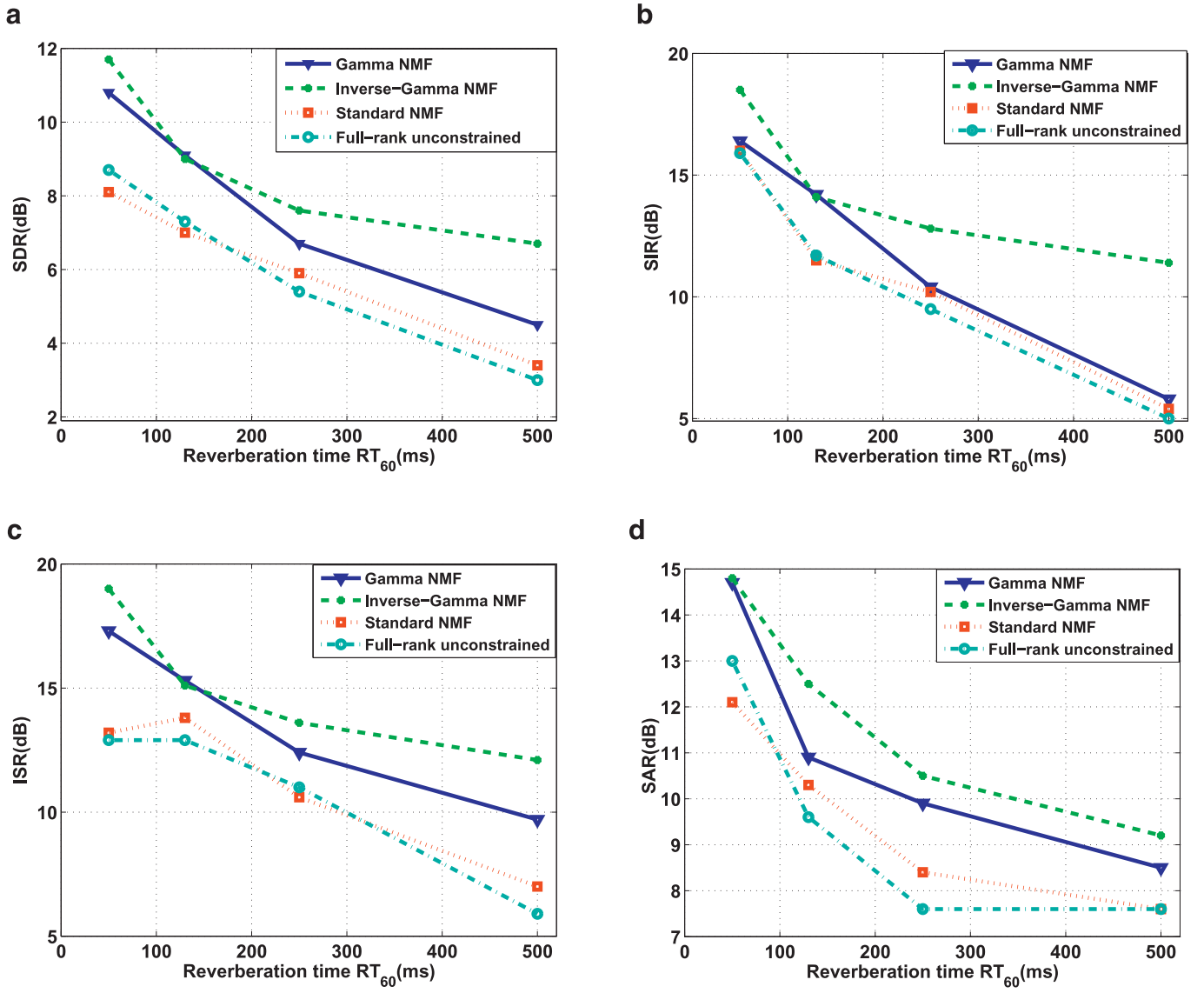

Fig. 1. Average blind source separation performance over stereo mixtures of three sources as a function of the reverberation time, measured in terms of (a) SDR, (b) SIR, (c) ISR, and (d) SAR.

The male and female speech signals and music signals are taken from the dev2 dataset of the SiSEC'08 "underdetermined speech and music mixtures" task (Vincent et al., 2009). Each source signal has a duration of 10 s. The sampling rate is equal to 16 kHz. The STFT frame size is 1024 with a frame shift of 512 samples. A sine window function is used at each frame to obtain the STFT coefficients. The number of components is set to 30 for NMF-based algorithms. 20 iterations of the EM algorithm is executed.

The separation quality is measured by calculating the BSS evaluation metrics; We use the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria expressed in decibels (dB), as defined in Vincent et al. (2007).

We have performed the experiments on both synthetically mixed and real-world data as explained in Sections 4.1 and 4.2 respectively.

4.1. Experiments on synthetic mixtures

The mixture signal is synthetically generated using the Roomsim Toolbox (Campbell et al., 2005) for a rectangular room of dimensions 6.25 × 3.75 × 2.5 m and omnidirectional microphones. Three sources are chosen for each mixture. There are 4 different mixture types: (1) Three male speech sources; (2) Three female speech sources; (3) Three nodrums music sources and (4) Three wdrums music sources. The nodrums data consists of three non-percussive sources while the wdrums includes the Drums musical instrument. The source directions w.r.t the microphones axis are taken as 30° , 70° and 120°. The source and microphone height are set to 1.1 m. The position of the microphone centers in (x,y) coordinates lies at (1.56 m, 1.87 m). The microphone spacing is set to $d = 5$ cm. We evaluate our proposed methods under different reverberant conditions corresponding to various RT_{60} parameters. The performance measures

Table 3
Average performance metrics obtained over speech mixtures.

RT_{60}	Algorithm	SDR (dB)	SIR (dB)	ISR (dB)	SAR (dB)
130 ms	Gamma Bayesian model	9.4	14	16.8	12.1
	Inverse-Gamma Bayesian model	7.8	12.8	16.1	11.8
250 ms	Gamma Bayesian model	7.3	12	13.6	10.6
	Inverse-Gamma Bayesian model	7.1	11.6	13.1	9.5

Table 4
Average performance metrics obtained over music mixtures.

RT_{60}	Algorithm	SDR (dB)	SIR (dB)	ISR (dB)	SAR (dB)
130 ms	Gamma Bayesian model	8.8	14.4	13.8	9.7
	Inverse-Gamma Bayesian model	10.2	15.4	14.1	13.2
250 ms	Gamma Bayesian model	6.1	8.8	11.2	9.2
	Inverse-Gamma Bayesian model	8.1	14	14.1	11.5

averaged over all sources and all mixtures are depicted against RT_{60} values in Fig. 1. As comparison materials, the results of the standard NMF (Arberet et al., 2010) and full-rank unconstrained model without NMF (Duong et al., 2010b), are also evaluated as indicated in the figure. It can be observed that our proposed Bayesian approaches outperform both state-of-the-art methods in all evaluation metrics. In the full-rank model of Duong et al. (2010b), since the model parameters are independently estimated at different frequencies, the well-known source permutation problem must be addressed. The DOA-based permutation alignment scheme was used for this purpose. However in the case of NMF and Bayesian NMF decomposition, the permutation problem can be avoided due to the joint estimation of the parameters.

The value chosen for the smoothing parameter ($\lambda = 0.5$), led to optimum performance in average for all signal types. For providing a quantitative analysis, we obtained the variance of the SDR metric evaluated for 60 λ values uniformly spaced between 0.3 and 0.9 in the case where the reverberation time is set to 250 ms. We observed that the standard deviation is equal to .13 dB and the best obtained metric corresponds to the case where $\lambda = 0.5$.

To measure where the superiority of each of two Bayesian frameworks stems from in modeling two audio mixture categories, music and speech, we calculated the average performance of the separation for speech and music mixtures separately as listed in Tables 3 and 4 respectively for the corresponding RT_{60} values equal to 130 ms and 250 ms. We observe that the Gamma Bayesian framework performs better for speech mixtures and the inverse-Gamma chain Bayesian model is better matched to the music signals. These results can be assigned to the more strict temporal continuity constraint imposed within applying the Gamma framework related to the inverse-Gamma model. Hence we may interpret that the strict temporal dependency constraint is better fitted to the speech signals in general. We have also observed that inserting the temporal dependencies into the model may not improve the separation quality for the percussive music sources because of their discontinuous nature. To illustrate this, we have analyzed the effect of reducing the temporal smoothness constraint in both Bayesian models and reported the obtained SDR for the wdrums mixture which contains percussive sources in Tables 5 and 6. RT_{60} is set to 250 ms. Reducing or eliminating the temporal continuity constraint in both models, can even lead to better performance for percussive sources "Drums" and "Hi-hat". However it can be observed that applying the temporal continuity constraint has improved the SDR corresponding to the "Bass" source.

Table 5
SDR metric (dB) for wdrums sources using the Gamma Bayesian model.

Smoothing parameter	Drums	Hi-hat	Bass
$\lambda = 0$	4.5	5.6	10.4
$\lambda = 0.5$	3.4	4.8	11.5

Table 6
SDR metric (dB) for wdrums sources using Inverse-Gamma Bayesian model.

Smoothing parameter	Drums	Hi-hat	Bass
$\alpha = 0.1$	5.1	7.7	11.0
$\alpha = 5$	4.4	7.2	12.2

The EM convergence is demonstrated in Fig. 2a and b for the two Bayesian frameworks representing the MAP criteria (Q_{1MAP} and Q_{3MAP} in the Appendices A.1 and B respectively) at each iteration. The MAP criterion is averaged over all mixtures.

4.2. Experiments on real-world mixtures

We perform this final experiment to compare the proposed Bayesian algorithms with state-of-the-art BSS algorithms submitted for evaluation to SiSEC 2008 over real-world mixtures of three or four speech sources. Two mixtures were recorded for each given number of sources, using either male or female speech signals. The room reverberation time was either 130 or 250 ms and the microphone spacing 5 cm Vincent et al. (2009). The average SDR achieved by each algorithm is listed in Table 7 for comparison. The SDR results of all algorithms were taken from Table III in Duong et al. (2010b). The Bayesian NMF-based methods outperform the other state-of-the-art methods. The better performance achieved using the proposed Bayesian approaches can be attributed to the choice of the prior structure as well as modeling the temporal smoothness. We have analyzed this by diminishing the smoothness constraint from both Bayesian models and compared the obtained average SDR values for real-world mixtures with the case where smoothness is considered.

For the Gamma model, we set λ to 0 to eliminate the smoothness. The obtained results are reported in Table 8. It can be implied that the temporal continuity constraint applied through the smoothing parameter λ, leads to improved performance in terms of average SDR metric.

For the second Inverse-Gamma model, we set the α parameter to 0.1 to reduce the sharpness of the prior distribution and thus alleviate the smoothness constraint. The results

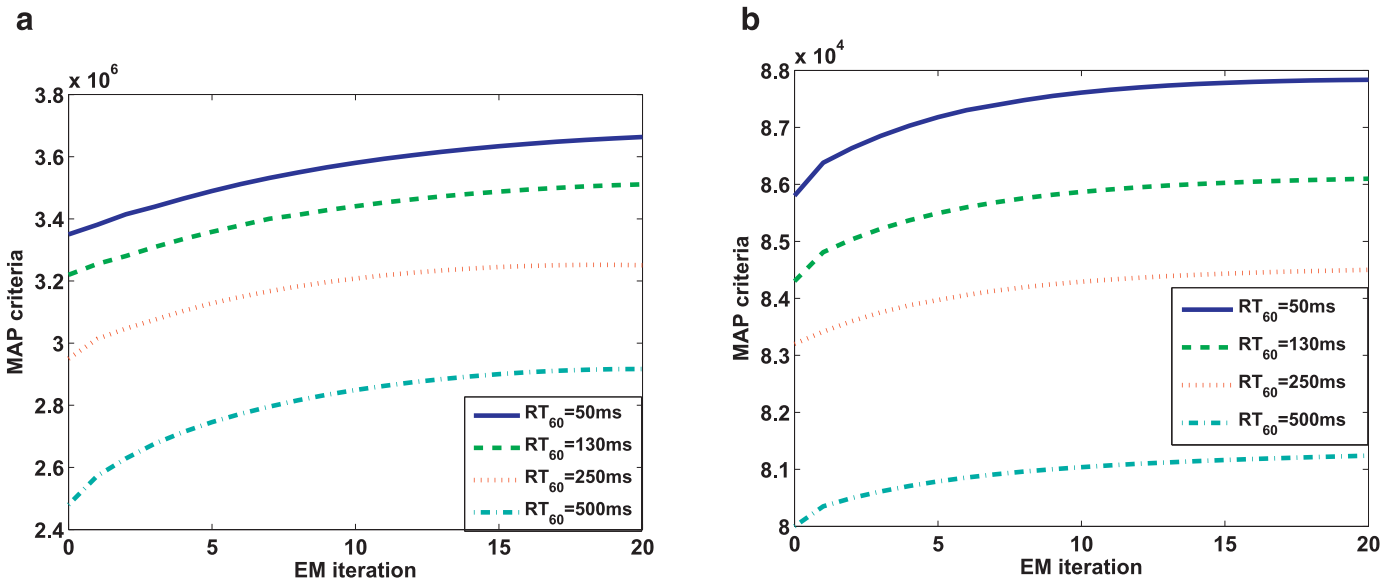

Fig. 2. EM convergence graphs for (a) Gamma Bayesian framework, (b) Inverse-Gamma chain Bayesian framework.

Table 7
Average SDR (dB) over the real-world test data of SiSEC 2008 with 5-cm microphone spacing.

RT_{60}	Algorithm	3 sources	4 sources
130 ms	Gamma Bayesian model	4.4	3.9
	Inverse-Gamma Bayesian model	4.6	3.4
	full-rank unconstrained model (Duong et al., 2010b)	3.3	2.8
	M.Cobes (Cobos and López, 2009)	2.3	2.1
	M.Mandel (Mandel and Ellis, 2007)	.1	−3.7
	R.Weiss (Weiss and Ellis, 2010)	2.9	2.3
	S.Araki (Araki et al., 2009)	2.9	–
	Z.El Chami (El Chami et al., 2008)	2.3	2.1
250 ms	Gamma Bayesian model	5.0	3.1
	Inverse-Gamma Bayesian model	4.9	3.5
	full-rank unconstrained model (Duong et al., 2010b)	3.8	2.0
	M.Cobes (Cobos and López, 2009)	2.2	1.0
	M.Mandel (Mandel and Ellis, 2007)	0.8	1.0
	R.Weiss (Weiss and Ellis, 2010)	2.3	1.5
	S.Araki (Araki et al., 2009)	3.7	–
	Z.El Chami (El Chami et al., 2008)	3.1	1.4

Table 8
Average SDR (dB) over the real-world test data for the Gamma Bayesian model.

RT_{60}	Algorithm	3 sources	4 sources
130 ms	Gamma Bayesian model with $\lambda = 0$	3.8	3.0
	Gamma Bayesian model with $\lambda = 0.5$	4.4	3.9
250 ms	Gamma Bayesian model with $\lambda = 0$	4.1	2.4
	Gamma Bayesian model with $\lambda = 0.5$	5.0	3.1

are reported in Table 9. Again, it can be observed that the smoothing constraint incorporated into the model through the inverse-Gamma prior hyperparameter α, leads to better average separation performance in terms of SDR.

Table 9
Average SDR (dB) over the real-world test data for the Inverse-Gamma Bayesian model.

RT_{60}	Algorithm	3 sources	4 sources
130 ms	Inverse-Gamma Bayesian model with $\alpha = 0.1$	3.6	2.9
	Inverse-Gamma Bayesian model with $\alpha = 5$	4.6	3.4
250 ms	Gamma Bayesian model with $\alpha = 0.1$	4	3.0
	Gamma Bayesian model with $\alpha = 5$	4.9	3.5

5. Conclusion

In this work, we proposed and developed two Bayesian NMF frameworks for separating stereo audio mixtures in a reverberant environment using a factorization of the source variance matrices in a full-rank model. The temporal dependencies of the audio signal are taken into account through incorporating two different prior distribution structures assumed for the activation coefficients: (1) a gamma prior distribution whose scale parameters are updated at each time frame and (2) an inverse-Gamma Markov chain prior distribution. The performance of the developed Bayesian methods are compared with each other as well as the standard NMF and the standard full-rank unconstrained modeling schemes by calculating the BSS evaluation metrics. It has been shown that the proposed Bayesian approaches outperform the previous state-of-the-art methods. It can also be concluded that the inverse-Gamma chain prior structure performs better for the music source separation and the Gamma prior structure with recursive updates is better fitted to the speech mixtures. Bayesian models offer both a strong theoretical framework and the possibility to manage constraints through models and priors.

Acknowledgment

This research was funded by the KU Leuven research grant GOA/14/005 (CAMETRON).

Appendix A

A.1. Derivation of the M-step update relations under Bayesian NMF framework with Gamma prior

MAP estimation of the temporal weights $h_{k,t}^{(n)}$ is equivalent to maximizing the cost function below:

$$\begin{aligned} Q_{1MAP} = &- \sum_{f,t,k,n} D_{KL}\left(\hat{\mathbf{R}}_{y_{n,k}}(f,t)|\mathbf{R}_{y_{n,k}}(f,t)\right) \\ &+ \left(a_{k,t}^{(n)}-1\right)\log h_{k,t}^{(n)} - \frac{h_{k,t}^{(n)}}{b_{k,t}^{(n)}} \\ f = 1:F, \quad t = 1:T, \quad k = 1:K, \quad n = 1:N \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} &\hat{R}_{y_{n,k}}(f,t) = \hat{Y}_{n,k}(f,t)\hat{Y}_{n,k}^{H}(f,t) + (\mathbf{I}-\mathbf{G}_{n,k}(f,t))\mathbf{R}_{y_{n,k}}(f,t) \\ &\hat{Y}_{n,k}(f,t) = \mathbf{G}_{n,k}(f,t)X(f,t) \\ &\mathbf{G}_{n,k}(f,t) = \mathbf{R}_{y_{n,k}}(f,t)(\mathbf{R}_X(f,t))^{-1} \\ &\mathbf{R}_{y_{n,k}}(f,t) = w_{f,k}^{(n)}h_{k,t}^{(n)}\mathbf{R}_n(f) \end{aligned} \tag{A.2}$$

The first term in (A.1) is equivalent to the Maximum Likelihood(ML) criterion as stated in Arberet et al. (2010). So Q_{1MAP} can be written as:

$$\begin{aligned} Q_{1MAP} = &\sum_{f,t,k,n} -\frac{1}{2}trace\left(\hat{\mathbf{R}}_{\mathbf{Y_n}}(f,t)\frac{\mathbf{R}_n^{-1}(f)}{w_{f,k}^{(n)}h_{k,t}^{(n)}}\right) \\ &+ \frac{1}{2}\log\det\left(\hat{\mathbf{R}}_{\mathbf{Y_n}}(f,t)\frac{\mathbf{R}_n^{-1}(f)}{w_{f,k}^{(n)}h_{k,t}^{(n)}}\right) + 1 \\ &+ \left(a_{k,t}^{(n)}-1\right)\log h_{k,t}^{(n)} - \frac{h_{k,t}^{(n)}}{b_{k,t}^{(n)}} - a_{k,t}^{(n)}\log\left(b_{k,t}^{(n)}\right) \end{aligned} \tag{A.3}$$

Then, the derivative of Q_{1MAP} w.r.t $h_{k,t}^{(n)}$ is obtained as:

$$\frac{dQ_{1MAP}}{dh_{k,t}^{(n)}} = \frac{\sum_f \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}} - \left(F - a_{k,t}^{(n)} + 1\right)h_{k,t}^{(n)} - \frac{\left(h_{k,t}^{(n)}\right)^2}{b_{k,t}^{(n)}}}{\left(h_{k,t}^{(n)}\right)^2} \tag{A.4}$$

Setting (A.4) to zero, will give us the update relation expressed in (14).

A.2. Initializing the parameters under Bayesian NMF framework with Gamma prior

We assume IS-NMF with Gamma prior model for the initialization step; Thus we should obtain the MAP estimation by minimizing Q_{2MAP} function:

$$Q_{2MAP} = \sum_{f,t,k,n} D_{IS}\left(v'_{k,f,t,n}|w_{f,k}^{(n)}h_{k,t}^{(n)}\right) - \log\left(P\left(h_{k,t}^{(n)}|a_{k,t}^{(n)}b_{k,t}^{(n)}\right)\right) \tag{A.5}$$

where $v'_{k,f,t,n} = |\frac{w_{f,k}^{(n)}h_{k,t}^{(n)}}{\sum_l w_{f,l}^{(n)}h_{l,t}^{(n)}}|^2 v_{f,t}^{(n)} + \frac{w_{f,k}^{(n)}h_{k,t}^{(n)}}{\sum_l w_{f,l}^{(n)}h_{l,t}^{(n)}}\sum_{l\neq k} w_{f,l}^{(n)}h_{l,t}^{(n)}$.

Therefore, the partial derivatives of (A.5) are obtained as:

$$\frac{dQ_{2MAP}}{dw_{f,k}^{(n)}} = \frac{T}{w_{f,k}^{(n)}} - \frac{1}{\left(w_{f,k}^{(n)}\right)^2}\sum_t \frac{v'_{k,f,t,n}}{h_{k,t}^{(n)}} \tag{A.6}$$

$$\frac{dQ_{2MAP}}{dh_{k,t}^{(n)}} = \frac{F}{h_{k,t}^{(n)}} - \frac{1}{\left(h_{k,t}^{(n)}\right)^2}\sum_f \frac{v'_{k,f,t,n}}{w_{f,k}^{(n)}} - \frac{a_{k,t}^{(n)}-1}{h_{k,t}^{(n)}} + \frac{1}{b_{k,t}^{(n)}} \tag{A.7}$$

where the first two terms on the right hand side of the equations are equal to the ML criterion(Q_{ML}) gradients. Therefore, the update relations are obtained as below:

$$w_{f,k}^{(n)} = \frac{1}{T}\sum_t \frac{v'_{k,f,t}}{h_{k,t}^{(n)}} \tag{A.8}$$

$$\begin{aligned} &h_{k,t}^{(n)} = \frac{\sqrt{p_1^2 - 4p_2p_0} - p_1}{2p_2} \\ &p_2 = \frac{1}{b_{k,t}^{(n)}} \quad p_1 = F - a_{k,t}^{(n)} + 1 \quad p_0 = -\sum_f \frac{v'_{k,f,t}}{w_{f,k}^{(n)}} \end{aligned} \tag{A.9}$$

The scale parameters $b_{k,t}^{(n)}$ of the prior are recursively updated at each time frame using (17).

Appendix B. Derivation of the M-step update relations under Bayesian NMF framework with inverse-gamma Markov chain prior

In the case of the Bayesian framework with Inverse-Gamma chain prior, the MAP criteria which should be maximized can be written as:

$$\begin{aligned} Q_{3MAP} = &\sum_{f,t,k,n} -\frac{1}{2}\mathrm{trace}\left(\hat{\mathbf{R}}_{\mathbf{Y_n}}(\mathrm{f},\mathrm{t})\frac{\mathbf{R}_{\mathrm{n}}^{-1}(\mathrm{f})}{\mathrm{w}_{\mathrm{f,k}}^{(\mathrm{n})}\mathrm{h}_{\mathrm{k,t}}^{(\mathrm{n})}}\right) \\ &+ \frac{1}{2}\log\left(\det\left(\hat{\mathbf{R}}_{\mathbf{Y_n}}(\mathrm{f},\mathrm{t})\frac{\mathbf{R}_{\mathrm{n}}^{-1}(\mathrm{f})}{\mathrm{w}_{\mathrm{f,k}}^{(\mathrm{n})}\mathrm{h}_{\mathrm{k,t}}^{(\mathrm{n})}}\right)\right) + 1 \\ &+ \log p\left(h_{k,t}^{(n)}|h_{k,t-1}^{(n)}\right) + \log p\left(h_{k,t+1}^{(n)}|h_{k,t}^{(n)}\right) \end{aligned} \tag{B.1}$$

So the gradient of (B.1) is obtained as follows:

$$\frac{dQ_{3MAP}}{dh_{k,t}^{(n)}} = \begin{cases} \dfrac{\sum_f \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}} + (\alpha - 1 - F)h_{k,t}^{(n)} - \frac{(\alpha+1)(h_{k,t}^{(n)})^2}{h_{k,t+1}^{(n)}}}{(h_{k,t}^{(n)})^2}, & t = 1 \\ \dfrac{\sum_f \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}} + (\alpha+1)h_{k,t-1}^{(n)} - (F+1)h_{k,t}^{(n)} - \frac{(\alpha+1)(h_{k,t}^{(n)})^2}{h_{k,t+1}^{(n)}}}{(h_{k,t}^{(n)})^2}, & 1 < t < T \\ \dfrac{\sum_f \frac{\hat{v}_{n,k}(f,t)}{w_{f,k}^{(n)}} + (\alpha+1)h_{k,t-1}^{(n)} - (\alpha+1+F)h_{k,t}^{(n)}}{(h_{k,t}^{(n)})^2}, & t = T \end{cases} \tag{B.2}$$

Zeroing the above gradient function, lead to the update equation stated in (19).

References

Araki, S., Nakatani, T., Sawada, H., Makino, S., 2009. Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem. In: Independent Component Analysis and Signal Separation. Springer, pp. 742–750.

Arberet, S., Ozerov, A., Duong, N.Q., Vincent, E., Gribonval, R., Bimbot, F., Vandergheynst, P., 2010. Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In: Proceedings of the 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA), 2010. IEEE, pp. 1–4.

Burred, J.J., Sikora, T., 2006. Comparison of frequency-warped representations for source separation of stereo mixtures. In: Audio Engineering Society Convention 121. Audio Engineering Society.

Campbell, D., Palomaki, K., Brown, G., 2005. A MATLAB simulation of "shoebox" room acoustics for use in research and teaching. Comput. Inf. Syst ICASSP 2010, 9–12.

Cobos, M., López, J., 2009. Blind separation of underdetermined speech mixtures based on DOA segmentation. IEEE Trans. Audio, Speech, Lang. Process.

Duong, N.Q., Vincent, E., Gribonval, R., 2009. Spatial covariance models for under-determined reverberant audio source separation. In: Applications of Signal Processing to Audio and Acoustics. IEEE, pp. 129–132.

Duong, N.Q., Vincent, E., Gribonval, R., 2010a. Under-determined convolutive blind source separation using spatial covariance models. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010. IEEE, pp. 9–12.

Duong, N.Q., Vincent, E., Gribonval, R., 2010b. Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. Audio, Speech, Lang. Process., 18 (7), 1830–1840.

Duong, N.Q., Vincent, E., Gribonval, R., 2010c. Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation. In: Latent Variable Analysis and Signal Separation. Springer, pp. 73–80.

Duong, N.Q., Vincent, E., Gribonval, R., 2013. Spatial location priors for gaussian model based reverberant audio source separation. EURASIP J. Adv. Signal Process. 2013, 1–11.

El Chami, Z., Pham, D., Serviere, C., Guerin, A., 2008. A new model based underdetermined source separation. In: Proceedings of the International on Workshop on Acoustic Echo and Noise Control (IWAENC), p. 147.

Essid, S., Févotte, C., 2013. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. IEEE Trans. Multimedia 415–425.

Févotte, C., 2011. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011. IEEE, pp. 1980–1983.

Févotte, C., Bertin, N., Durrieu, J.-L., 2009. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural computation, 793–830.

Févotte, C., Le Roux, J., Hershey, J.R., 2013. Non-negative dynamical system with application to speech and audio. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE, pp. 3158–3162.

Kim, M., Smaragdis, P., 2013. Single channel source separation using smooth nonnegative matrix factorization with Markov random fields. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2013. IEEE, pp. 1–6.

Mandel, M.I., Ellis, D.P., 2007. EM localization and separation using interaural level and phase cues. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007. IEEE, pp. 275–278.

Mohammadiha, N., Smaragdis, P., Leijon, A., 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio, Speech, Lang. Process. 2140–2151.

Mohammadiha, N., Taghia, J., Leijon, A., 2012. Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012. IEEE, pp. 4561–4564.

Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., Sagayama, S., 2010. Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms. In: Latent Variable Analysis and Signal Separation. Springer, pp. 149–156.

Nikunen, J., Virtanen, T., 2014. Direction of arrival based spatial covariance model for blind sound source separation. IEEE/ACM Trans. Audio, Speech Lang. Process. 22 (3), 727–739.

Ozerov, A., Févotte, C., 2010. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio, Speech, Lang. Process. 18 (3), 550–563.

Sawada, H., Kameoka, H., Araki, S., Ueda, N., 2013. Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans. Audio, Speech Lang. Process. 21 (5), 971–982.

Smaragdis, P., Fevotte, C., Mysore, G., Mohammadiha, N., Hoffman, M., 2014. Static and dynamic source separation using nonnegative factorizations: a unified view. Signal Process. Mag., IEEE, 66–75.

Vincent, E., 2006. Musical source separation using time-frequency source priors. Audio, Speech, Lang. Process., IEEE Trans. 91–98.

Vincent, E., Araki, S., Bofill, P., 2009. The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation. In: Independent Component Analysis and Signal Separation. Springer, pp. 734–741.

Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P., 2007. First stereo audio source separation evaluation campaign: data, algorithms and results. In: Independent Component Analysis and Signal Separation. Springer, pp. 552–559.

Virtanen, T., 2007. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Lang. Process. 15 (3), 1066–1074.

Weiss, R.J., Ellis, D.P., 2010. Speech separation using speaker-adapted eigenvoice speech models. Comput. Speech Lang. 16–29.

Wilson, K.W., Raj, B., Smaragdis, P., 2008. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In: Interspeech, pp. 411–414.

Winter, S., Kellermann, W., Sawada, H., Makino, S., 2007. Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization. EURASIP J. Appl. Signal Process 2007 (1), 81–92.

Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52 (7), 1830–1847.